Grounded Action Transformation for Robot Learning in Simulation

JOSIAH HANNA AND PETER STONE
The University of Texas at Austin
Austin, TX 78712 USA
{jphanna,pstone}@cs.utexas.edu

Abstract

- Robot learning in simulation is a promising alternative to the sample cost of real world learning.
- Policies learned in simulation often perform worse than hand-coded policies on the physical robot.
- We propose the **Grounded Action Transformation** algorithm for robot learning in simulation.
  - Our approach results in a **43.27% improvement** in humanoid bipedal walk velocity compared to a state-of-the-art hand-coded walk.

Problem Definition

Environment $E = \langle \mathcal{S}, \mathcal{A}, c, P \rangle$

- Robot in state $s \in \mathcal{S}$ chooses action $a \in \mathcal{A}$ according to policy π.
- Environment, E, responds with a new state $S_{t+1} \sim P(\cdot|s,a)$.
- Cost function c defines a scalar cost for each (s,a).
- Policy performance measured by expected sum of costs:

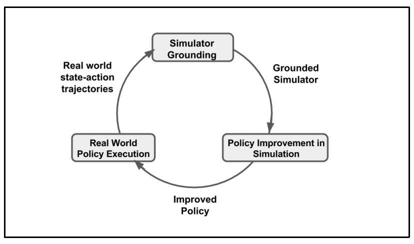
$$J(\pi) := \mathbb{E}_{S_1,A_1,\ldots,S_L,A_L}\left[\sum_{t=1}^{L} c(S_t, A_t)\right]$$

Simulator $E_{\texttt{sim}} = \langle \mathcal{S}, \mathcal{A}, c, P_{\texttt{sim}} \rangle$.

- Identical to E but different transition probabilities.

Goal: Minimize $J_{\texttt{sim}}(\pi)$ such that $J(\pi)$ also decreases.

Grounded Simulation Learning [1]



1. **Collect** sample trajectories with initial policy on physical robot.
2. **Ground** simulation such that the initial policy produces similar trajectories in simulation.
3. **Optimize** the policy **in simulation** to find better policy parameters.
4. Set the new policy to be the initial policy and repeat.

Grounding Simulation to Reality

- P_ϕ: Simulator dynamics $P_{\texttt{sim}}$ with parameters ϕ.
- Given:
  - $\mathcal{D}$: a data-set of real world state-action trajectories.
  - d: a measure of similarity between probability distributions.

Grounding simulation means finding simulation parameters $\phi^\star$ such that:

$$\phi^\star = \operatorname*{argmin}_{\phi} \sum_{(S_t,A_t)\in\mathcal{D}} d\left(P(\cdot|S_t,A_t), P_\phi(\cdot|S_t,A_t)\right)$$

Grounded Action Transformation

Augment simulation with an **action transformation module**:

- Replace robot's action $\mathbf{a}_t$ with an action that produces a more **realistic** transition.
- Learn this action as a function $g(\mathbf{s}_t, \mathbf{a}_t)$.

g composed of two functions:

- Robot's dynamics: $f: \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$
- Simulator's inverse dynamics: $f^{-1}_{\texttt{sim}}: \mathcal{S} \times \mathcal{S} \rightarrow \mathcal{A}$.

Replace robot's action $\mathbf{a}_t$ with $\hat{\mathbf{a}}_t := f^{-1}_{\texttt{sim}}(\mathbf{s}_t, f(\mathbf{s}_t, \mathbf{a}_t))$.

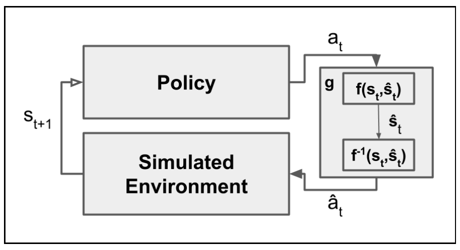


Figure 1: The augmented simulator induced by GAT.

GAT Training Procedure

f and $f^{-1}_{\texttt{sim}}$ trained with **supervised learning**.

- Record sequence $S_t, A_t, \ldots$ on robot and in simulation.
- Supervised learning of g:
  - $f^{-1}_{\texttt{sim}}: (S_t, A_t) \rightarrow S_{t+1}$
  - $f: (S_t, S_{t+1}) \rightarrow A_t$
- **Neural networks** in this work.

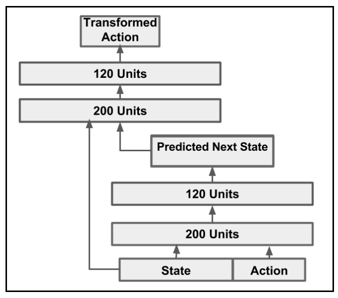


Figure 2: Neural network architecture used to learn GAT's action modification function.

Empirical Study

Applied GAT to learning **fast, bipedal walks** for the **SoftBank NAO** robot.

- Task: Walk forward towards a target.
- Initial policy: University of New South Wales Walk Engine [3].
- Policy optimization with **CMA-ES** stochastic search method [2].

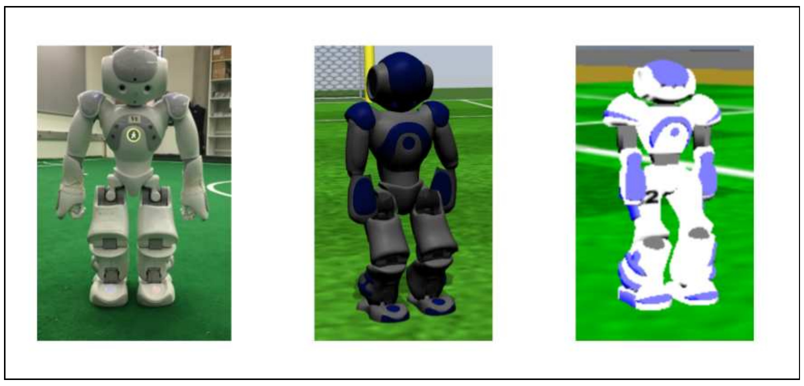


Figure 3: Walk policies learned within the **Gazebo Simulator** (center) and **SimSpark Simulator** (right) were successfully transferred to the **SoftBank NAO** robot (left). Walk policies learned within **SimSpark** were successfully transferred to the **Gazebo** simulator.

Simulation to Nao:

Method	Velocity (cm/s)	% Improve
Initial policy	19.52	0.0
SimSpark, first iteration	26.27	34.58
SimSpark, second iteration	27.97	43.27
Gazebo, first iteration	26.89	37.76

SimSpark to Gazebo:

Method	% Improve	Failures
GAT	**22.48**	**1**
No Ground	11.094	7
Noise-Envelope	18.93	5

Discussion

- Demonstrated GAT can **optimize** policies in simulation and **transfer** them to physical robots.
- GAT treats simulator as a black-box — requiring no special knowledge of how to modify simulation.

Future Work

- Extending to other robotics tasks and platforms (e.g., manipulation with contacts).
- Characterizing when grounding actions works and when does it not.

[1] A. Farchy, S. Barrett, P. MacAlpine, and P. Stone. Humanoid robots learning to walk faster: From the real world to simulation and back. In *Twelfth International Conference on Autonomous Agents and Multiagent Systems*, 2013.

[2] N. Hansen. The cma evolution strategy: A tutorial. 2011.

[3] B. Hengst, M. Lange, and B. White. Learning to control a biped with feet. In *Humanoids*, 2011.

Acknowledgments

This work has taken place in the Learning Agents Research Group (LARG) at the Artificial Intelligence Laboratory, The University of Texas at Austin. LARG research is supported in part by NSF (CNS-1330072, CNS-1305287, IIS-1637736, IIS-1651089), ONR (21C184-01), and AFOSR (FA9550-14-1-0087). Josiah Hanna is supported by an NSF Graduate Research Fellowship. Peter Stone serves on the Board of Directors of, Cogitai, Inc. The terms of this arrangement have been reviewed and approved by the University of Texas at Austin in accordance with its policy on objectivity in research.